SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

19IT601– Data Science and Analytics

III Year / VI Semester

Unit 3 – PREDICTIVE MODELING AND MACHINE LEARNING

Topic 8: Unsupervised Learning

Unsupervised Learning

Unsupervised Learning Algorithms

Clustering
Clustering techniques are unsupervised in the sense that the data scientist does not determine, in advance, the labels to apply to the clusters.

The structure of the data describes the objects of interest and determines how best to group the objects. Clustering is a method often used for exploratory analysis of the data.

In clustering, there are no predictions made. Rather, clustering methods find the similarities between objects according to the object attributes and group the similar objects into clusters.

Clustering techniques are utilized in marketing, economics, and various branches of science.

Unsupervised Learning

There are primarily two categories of clustering:

- Hierarchical clustering
- Partitioning clustering

Hierarchical clustering is further subdivided into:

- Agglomerative clustering
- Divisive clustering

Partitioning clustering is further subdivided into:

- K-Means clustering
- Fuzzy C-Means clustering

Unsupervised Learning

K-Means Clustering

Given a collection of objects each with n measurable attributes and a chosen value k of the number of clusters, the algorithm identifies the k clusters of objects based on the objects proximity to the centers of the k groups.

The algorithm is iterative with the centers adjusted to the mean of each cluster's n-dimensional vector of attributes

K-means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster.

In K-Means, each cluster is associated with a centroid. The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid.

Unsupervised Learning

K-Means Algorithm

1.Choose the value of k (i.e number of clusters) and the initial guesses for the centroids

2. Compute the distance from each data point to each centroid, and assign each point to the closest centroid

3.Compute the centroid of each newly defined cluster from step 2

4.Repeat steps 2 and 3 until the algorithm converges (no changes occur)

Stopping Criteria for K-Means Clustering

There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:

- Centroids of newly formed clusters do not change
- Points remain in the same cluster
- Maximum number of iterations are reached

Unsupervised Learning

Distance Measure

Distance measure determines the similarity between two elements and influences the shape of clusters.

K-Means clustering supports various kinds of distance measures, such as:

- Euclidean distance measure
- Manhattan distance measure
- A squared euclidean distance measure
- Cosine distance measure

Unsupervised Learning

Determining the Number of Clusters

- The value of k can be chosen based on a reasonable guess or some predefined requirement.

- However, even then, it would be good to know how much better or worse having k clusters versus k – 1 or k + 1 clusters would be in explaining the structure of the data.

- Next, a heuristic using the Within Sum of Squares (WSS) metric is examined to determine a reasonably optimal value of k.

- In other words, WSS is the sum of the squares of the distances between each data point and the closest centroid.

- The term indicates the closest centroid that is associated with the ith point.

- If the points are relatively close to their respective centroids, the WSS is relatively small.

Unsupervised Learning

Application of K-Means

Some specific applications of k-means are

- image processing,
- medical, and
- customer segmentation

THANK YOU